Week 4 - Assignment

Programming for Data Science 2024

Exercises for the topics covered in the fourth lecture.

The exercise will be marked as passed if you get **at least 10/15** points.

Exercises must be handed in via **ILIAS** (Homework assignments). Deliver your submission as a compressed file (zip) containing one .py or .ipynb file with all exercises. The name of both the .zip and the .py/.ipynb file **must** be *SurnameName* of the two members of the group. Example: Riccardo Cusinato + Athina Tzovara = *CusinatoRiccardo_TzovaraAthina.zip* .

It's important to use comments to explain your code and show that you're able to take ownership of the exercises and discuss them.

You are not expected to collaborate outside of the group on exercises and submitting other groups' code as your own will result in 0 points.

For questions contact: *riccardo.cusinato@unibe.ch* with the subject: *Programming for Data Science 2024*.

Deadline: 14:00, March 21, 2024.

Exercise 1 - Create Dataframes 5 points

Create a DataFrame *episodes_df* with the columns **ses**, **ep**, and **title**, as below:

ses	ep	title
1	1	One
1	2	Two
2	1	Three
2	2	Four

Create a DataFrame *imdb_df* with the columns **ses**, **ep**, and **score**, as below:

ses	ep	score
1	1	8.4
1	2	8.1
2	1	7.9
2	2	7.7

Merge the two DataFrames. Then, find and print the title of the episode with the higest score.

NB: To merge the two dataframes you have to use the *merge* method:

```
merged_df = episodes_df.merge(imdb_df, on=['ses', 'ep'])
```

1. By manipulating the dataframes, find and print the title of the episode with the higest score. (*3 points*)

In [ ]:
```
###
# YOUR CODE HERE
###
```

2. Change the **score** of the entry with the title "Three" in the DataFrame you created in Task 1 and print the result. The new score should be 6. (*2 points*)

In [ ]:
```
###
# YOUR CODE HERE
###
```

Exercise 2 - Load DataFrames 6 points

1. Load the two CSV files 'silicon_valley_episodes.csv' and 'silicon_valley_imdb.csv', found in the "data" directory, as DataFrames. Merge the two DataFrames as in the first task, using **season** and **episode_num** as keys to merge on. (*2 points*)

```
episodes_df = pd.read_csv('./data/silicon_valley_episodes.csv')
imdb_df = pd.read_csv('./data/silicon_valley_imdb.csv')
```

In [ ]:
```
###
# YOUR CODE HERE
###
```

2. Create a new DataFrame, called **df_best**, containing only the episodes with an **imdb_rating** at or above 9. Use the DataFrame created in the previous task as a starting point. (*2 points*)

In [ ]:
```
###
# YOUR CODE HERE
###
```

3. Find mean number of **us_viewers** for episodes with an IMDB score greater than or equal to 9, and for episodes with an IMDB score lower than 9, and print the means. (*2 points*)

In [ ]:
```
###
# YOUR CODE HERE
###
```

Exercise 3 - DataFrames Ufuncs 4 points

Create the two dataframe *df1* and *df2* with the following code:

In [ ]:
```
import numpy as np
import pandas as pd

df1 = pd.DataFrame(
    np.arange(1, 10).reshape(3, 3),
    columns=["a", "b", "c"],
    index=["1", "2", "3"]
)

df2 = pd.DataFrame(
    np.arange(1, 10).reshape(3, 3) / 2,
    columns=["a", "b", "d"],
    index=["1", "2", "4"]
)
```

1. Add the two dataframes together, with the appropriate pandas method, and print the result. (*0.5 points*)

In [ ]:
```
###
# YOUR CODE HERE
###
```

2. Add the underlying numpy objects of the two dataframes, and print the result. (*0.5 points*)

In [ ]:
```
###
# YOUR CODE HERE
###
```

3. Compare the two results that you obtained and comment if and **why** they are different. (*3 points*)

In [ ]:
```
###
# YOUR COMMENT HERE
###
```